Reinforcement Guided Multi-Task Learning Framework for Low-Resource Stereotype Detection

Rajkumar Pujari[1] Erik Oveson[2] Priyanka Kulkari[2] Elnaz Nouri[3]

[1]Purdue University, [2]Microsoft, Redmond, [3]Microsoft Research, Redmond

rpujari@purdue.edu, {erikov,priyak,elnouri}@microsoft.com

Scan Me

Microsoft

Abstract

- [1] show that there are significant reliability issues with the existing '*Stereotype Detection*' datasets. We annotate a focused evaluation set for '*Stereotype Detection*' task that addresses those pitfalls by de-constructing various ways in which stereotypes manifest in text.
- We propose a reinforcement-learning agent that guides a multi-task learning model by learning to identify the training examples from neighboring tasks (hate speech detection, offensive language detection, misogyny detection, etc.,) that help the target task ('*Stereotype Detection*'). We show that the proposed models achieve significant empirical gains over existing baselines on all the tasks.

Motivation

- Empirical success of large Pretrained Language Models (PLMs) led to them being ubiquitously used in daily-life applications that interact with humans. Unsupervised training on huge, un-curated datasets results in harmful text and societal text creeping in their outputs
- This motivates a two-pronged solution:
  1) To diagnose and de-noise the bias in the PLMs
  2) **To identify & regulate harmful text externally at the output**
- This work focuses on the task of *identifying stereotypical associations* in text. *Stereotypes* differ from other harmful text such as hate speech, misogyny, abuse, threat, insult etc., in two important ways:
  1) They could also express a positive sentiment towards the target
  2) We need knowledge of their existence in the society to identify them

Our Dataset

- [1] demonstrate that existing datasets suffer from conceptual and operational issues. Diagnostic datasets, by nature, also suffer from lack of coverage of subtle manifestations of stereotypes in text.
- We address the coverage issue by collecting data samples for annotation from two subreddits: /r/Jokes (stereotype-rich) and /r/AskHistorians (stereotype-poor)
- To avoid operational and conceptual pitfalls, we ask the annotators to answer *three* questions for each sample:
  1) Is an over-simplified belief about a type of person "intentionally" expressed?
  2) Is there an"unintentional", widely-known stereotypical association present?
  3) Does the sentence seem made up (unlikely to occur in regular discourse)?
- Examples of data categories in our dataset:
  1) Ethiopians like stew (*explicit stereotype*)
  2) The lawyer misrepresented the situation and tricked the person (*implicit stereotype*)
  3) Jews spend money lavishly (*anti-stereotype*)
  4) There is an Asian family that lives down the street (*non-stereotype*)

Data Type	Size
Explicit Stereotypes	742
Implicit Stereotypes	282
Non-Stereotypes	1197

Figure: Statistics of Our Dataset

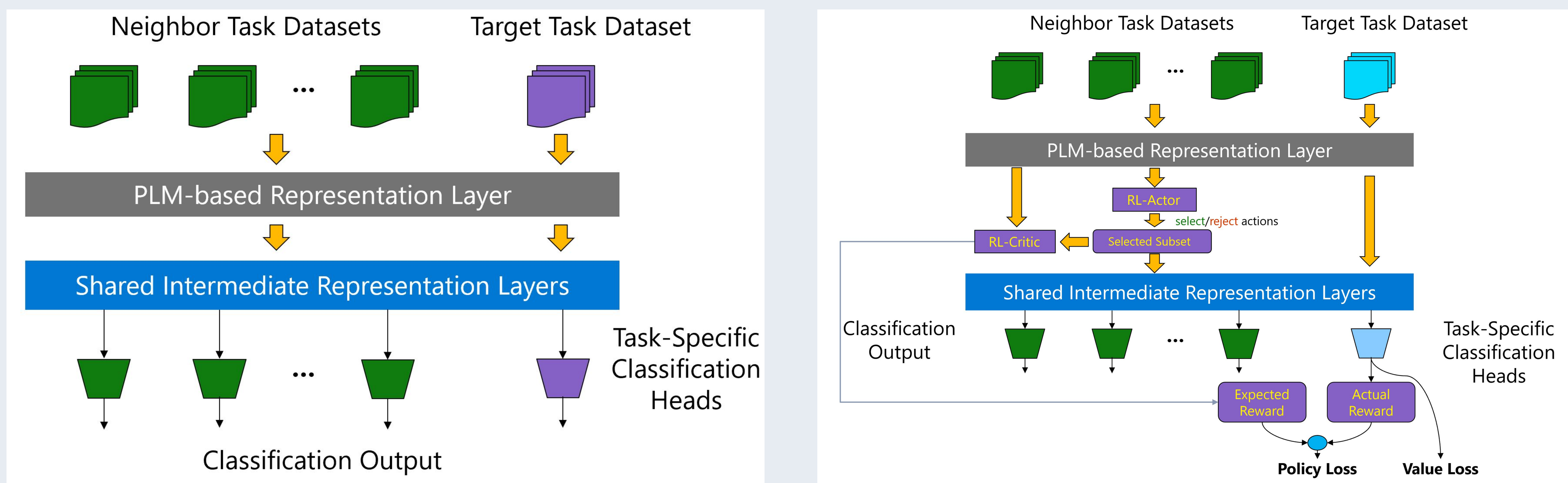
Multi-Task Learning Model

- Several datasets for harmful language identification such as hate speech detection, offensive language detection, misogyny detection and toxicity detection are widely available. They often contain overlapping objectives. For example:
  1) She may or may not be a jew but, she's certainly cheap! (insult, stereotype)
  2) Burn in hell, you Asian bastard! (abuse, stereotype)
- We hypothesize that solving these tasks require understanding largely similar linguistic characteristics of the text. We call these tasks "neighbor tasks".
- As the tasks have "overlapping objective" and require "understanding similar linguistic characteristic" of text, leveraging the intermediate representations from the neighbor tasks should benefit the target task.

Reinforcement-Guided MTL Model

- The main intuition behind the RL-MTL model is that "*not all examples from the neighbor task are equally useful in learning the target task*".
- We train an RL-agent on top of the MTL model to identify examples from neighbor tasks, which are beneficial for the target task
- Algorithm to train the RL agent:
  Step 1: For each example in neighbor task, RL-actor makes a select/reject decision
  Step 2: MTL model is trained on the selected examples
  Step 3: The RL-actor is assigned a reward based on the change in the performance on the target task
  Step 4: The loss between RL-actor's actual reward and RL-critic's expected reward is used to train the RL-agent

Model Architecture



(a) Architecture of the Multi-Task Learning Model

(b) Architecure of the RL-Guided MTL Model

Experiments

- We perform experiments using *six* datasets in *three* phases:
  Phase 1: Fine-tune PLM-based classifier
  Phase 2: Train a multi-task learning (MTL) model for all the datasets
  Phase 3: Train RL-guided MTL model for each task as target task
- We experiment with four PLMs as base-classifiers: BERT-base, BERT-large, BART-large and XLNet-large
- We use the following datasets for our experiments:
  1) Hate Speech Detection (de Gilbert et al., 2018)
  2) Offensive Language Detection (Davidson et al., 2017)
  3) Misogyny Detection (Fersini et al., 2018)
  4) Coarse-Grained Stereotype Detection (combination of StereoSet and CrowS-Pairs)
  5) Fine-Grained Stereotype Detection (our dataset)
  6) Jigsaw Toxicity Dataset (used only for training)

Results

Model	Hate Speech Detection	Offense Detection	Misogyny Detection	Coarse-grained Stereotypes	Fine-grained Stereotypes
BERT-base	66.47	66.13	74.16	65.71	61.36
BERT-large	67.05	63.90	72.13	59.63	55.42
BART-large	68.91	65.86	73.12	63.40	54.64
XLNet-large	59.14	48.33	63.16	63.71	53.80
Multi-Task Learning					
BERT-base + MTL	69.21	68.57	73.48	68.29	65.00
BERT-large + MTL	69.78	65.14	73.94	61.96	61.65
BART-large + MTL	67.79	68.03	74.40	65.77	64.90
XLNet-large + MTL	61.68	46.35	64.42	65.21	57.00
RL-guided Multi-Task Learning					
BERT-base + RL-MTL	**72.06**	**68.97**	74.48	**74.18**	65.72
BERT-large + RL-MTL	69.82	65.97	**75.21**	70.88	64.74
BART-large + RL-MTL	69.60	66.76	75.14	74.11	**67.94**
XLNet-large + RL-MTL	61.97	47.60	63.21	67.98	56.37

Figure: Results on all the Datasets for various phases. Macro-F1 score has been reported.

Impact of MTL Prior on RL-MTL

- In our experiments, we initialize RL-MTL model with trained parameters from the MTL model. In this ablation, we initialize the RL-MTL model randomly and observe the difference in performance.

Task	MTL Initialization	Random Initialization
Hate Speech Detection	72.06	70.23
Offense Detection	68.97	67.23
Misogyny Detection	74.78	71.10
Coarse-grained Stereotypes	74.18	60.42
Fine-grained Stereotypes	65.72	57.32

Figure: Macro-F1 scores on each task with 1) MTL initialization and 2) random initialization for the RL-Guided MTL model

Neighbor Task Impact

- We study the impact of each neighbor task with each task as a target task

Target \ Neighbor	Hate Speech Detection	Offense Detection	Misogyny Detection	Coarse-grained Stereotype
Hate Speech	-	69.69	70.07	**71.10**
Offensive Language	66.71	-	66.56	**67.39**
Misogyny	70.98	**75.87**	-	73.89
Coarse Stereotype	66.15	**67.40**	63.82	-
Fine Stereotype	**63.80**	63.65	59.94	56.12

Figure: Macro-F1 scores on each Target Task for each individual Neighbor Task.

Conclusion

- We tackle the problem of *Stereotype Detection* from *data annotation* and *low-resource computational framework* perspectives
- We devise a *focused annotation task* in conjunction with selective data candidate collection to create a fine-grained evaluation set for the task
- We utilize neighbor tasks with abundance of high-quality gold data in our *multi-task learning model*. We further propose an *RL-guided multi-task learning model* that learns to select examples from the neighbor tasks which benefit the target task.

References

[1] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach.
Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets.
In *ACL-IJCNLP 2021*, August 2021.

Resources

https://github.com/pujari-rajkumar/rl-guided-multitask-learning